DOCUMENT RESUME

ED 086 740 TM 003 400

AUTHOR Ebel, Robert L.
TITLE Can Classroom Teachers Write Good True-False Test Items
PUB DATE 73
NOTE 8p.; Paper presented at the 1974 Convention of the National Council on Measurement in Education, Chicago, Illinois

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Guessing (Tests); *Multiple Choice Tests; Scoring; *Teachers; *Test Construction; Testing Problems; Test Interpretation; *Test Reliability; *Test Validity
IDENTIFIERS *True False Tests

ABSTRACT

True-false achievement test items written by typical classroom teachers show about two-thirds of the discrimination of their multiple-choice test items. This is about what should be expected in view of the higher probability of chance success on the true-false items. However, at least half again as many true-false items as multiple-choice items can be answered comfortably in the same period of time. Thus the larger number of true-false items compensates for the lower discriminating power of the individual items. The data of this study support the belief that in the hands of typical classroom teachers the two item forms can be expected to give approximately equal reliabilities for tests of equal duration. (Author)

FILMED FROM BEST AVAILABLE COPY

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 086740

TM 003 400

Abstract

Can Classroom Teachers Write Good True-False Test Items?

Robert L. Ebel
Michigan State University

True-false achievement test items written by typical classroom teachers show about two-thirds of the discrimination of their multiple-choice test items. This is about what should be expected in view of the higher probability of chance success on the true-false items. However, at least half again as many true-false items as multiple-choice items can be answered comfortably in the same period of time. Thus the larger number of true-false items compensates for the lower discriminating power of the individual items.

The data of this study support the belief that in the hands of typical classroom teachers the two item forms can be expected to give approximately equal reliabilities for tests of equal duration.

Presented at the 1974 Convention of the National Council on Measurement in Education, Chicago, Illinois

Can Classroom Teachers Write Good True-False Test Items

1. Objective of the inquiry

True-false test items are regarded with disfavor by some test specialists and test users. They are suspected of being often trivial or ambiguous and always susceptibile to guessing. Ebel [1] has argued that these faults are not inherent in the form, and need not seriously limit its usefulness. He has provided a rationale for the validity of the form in tests of educational achievement, and has shown that highly reliable test scores can be obtained from true-false tests. However, it has been suggested that effective use of the form requires special talent, and that typical classroom teachers are unlikely to be able to use it effectively.[2,3] The present study was designed to shed some light on this question.

2. Data source and method

The source of the data was an item writing exercise used in a course in classroom testing at Michigan State University. After study of some basic principles of item writing the students are asked to test their skill. They find or create two short passages presenting important ideas which most of their classmates are unlikely to know. These two passages provide the background for two test items; one true-false and one multiple choice. The items are intended to discriminate those who know from those who do not know the ideas presented in the paragraphs. Sample background paragraphs and test items are shown in Exhibit 1.

The discriminating power of each item is determined in this way. Each student reads, or circulates a copy of his items to eight or nine of his classmates. They pick what seems to them the best answer. Then the student reads

Exhibit 1

Items for Discrimination Try-out

1. True-false item

According to the law of averages, if the first ten tosses of a coin give 8 heads and only 2 tails, the next ten can be expected to give more tails than heads. (T or F)

Background Paragraph

As applied to the tossing of a coin, the law of averages indicates that the proportion of heads can be expected to approach 50% more and more closely as the number of tosses becomes larger and larger. But this can happen even if, as is also likely, the difference between the number of heads and tails tends to get larger as the number of tosses gets larger. Thus the law of averages does not require that an excess of heads early in the series of tosses be offset by an excess of tails later in the series. However disproportionate the outcome of the early tosses, it could hardly have any influence on the outcome of independent tosses made later.

2. Multiple-choice item

How are members of the armed forces handled in the compilation of employment statistics?

1. as employed
2. as unemployed
3. as unemployable
4. as not in the labor force

Background Paragraph

Employment statistics are collected monthly by trained interviewers who obtain information from approximately 50,000 households each month. Enough information is obtained to classify persons 16 years of age and over as (1) employed (2) unemployed or (3) not in the labor force. Members of the armed forces are considered to be not in the labor force.

or circulates a copy of the background information on which the item was based. Finally the items are read or circulated to the classmates a second time. The difference between the number of correct answers before and after information is used as the measure of item discrimination.

3. Results of the study

The data in Table 1 summarize results obtained from the true-false and multiple-choice items written by about 250 practicing teachers or prospective teachers in five classes at Michigan State University in 1972 and 1973. Figures in the first column of the table are post-test minus pre-test differences in number of correct responses from the tryout group. Maximum size of any tryout group was 10 students (nine in addition to the item writer).

Numbers in the second and third columns of the table show the numbers of true-false and multiple-choice items that showed each of the indicated levels of discrimination. That is, there were no true-false items and eleven multiple-choice items which no one answered correctly on the pre-test but which all nine students answered correctly on the post-test.

The three rows at the bottom of the table summarize the column data. The first row, shows how many items of each type were written. The second row, gives the total post-test minus pre-test difference for all items of each kind. The third row, obtained by dividing the second by the first, gives the mean difference for all items of that type.

The overall mean post-test minus pre-test difference for the 247 true-false items written in these five classes was 2.63. The corresponding mean for the multiple-choice items was 4.17.

Table 1

Discrimination* of Items Written by Teachers

Post-Pre Difference	True-False Items	Multiple-Choice Items
9	0	11
8	3	15
7	6	22
6	22	34
5	22	34
4	25	38
3	50	26
2	38	20
1	41	22
0	25	17
-1	9	4
-2	3	1
-3	2	2
-4	1	0
Number	247	246
Sum	650	1027
Mean	2.63	4.17

* Number of correct answers with information minus number of correct answers without information.

4. Theoretical analysis

It is instructive to consider at this point what relative discriminating power (i.e. post-test minus pre-test difference) it is reasonable to expect from the two kinds of items. Table 2 presents figures relative to this question.

Table 2

Theoretical Maximum Discriminating Power

Proportion correct	True-false	Multiple-choice
Post-test (ideal)	1.00	1.00
Pre-test (chance)	.50	.25
Available for discrimination	.50	.75

A student guessing blindly on pre-test true-false items could expect to answer half of them correctly. With ideal students responding to ideal true-false test items, all of them would give correct answers on the post-test. Hence the discrimination that it is reasonable to expect of a perfect true-false item is .50. By similar reasoning it is reasonable to expect a discrimination of .75 for an ideal multiple-choice item to which ideal students are responding. (An ideal student, in this context, is totally ignorant of the subject of the item on the pre-test, and totally informed on the post-test). If the true-false items written by teachers fall as far short of perfection, proportionally, as do their multiple-choice items, one would expect the ratio of discriminations to be 2 to 3 or .67. The ratio obtained in this study was .63.

5. Conclusions and implications

Our sample of classroom teachers did better in writing multiple-choice

test items than in writing true-false test items. Item for item, their multiple-choice items are clearly more discriminating. But, this is hardly a fair comparison, since true-false items can be written more quickly by teachers, and responded to more quickly by students, than multiple-choice test items.

As has been shown in another paper,[4] there is a close relation between the mean index of discrimination for the items in a test of specified length and the reliability of that test. While the indices of discrimination obtained for the items written by teachers in this study are obviously not exactly the same as indices of discrimination obtained from upper-lower 27% groups, they would seem to be closely analogous in meaning.

Thus it seems very likely that a typical teacher can measure achievement as reliably with true-false as with multiple-choice items provided that about five true-false items are used instead of three multiple-choice items. Most teachers can probably write five true-false items in the time required to write three multiple-choice items, and most students can probably answer them comfortably in about the same ratio. Hence it seems that there is no very sound basis for any recommendation that classroom teachers give preference to multiple-choice over true-false test items.

References

1. Ebel, Robert L. "The Case for True-False Test Items." School Review, 78: 373-389; May, 1970.

2. Storey, A. G. "Review of Evidence or the Case Against the True-False Item," J. Ed. Res., 59: 282-5; 1966.

3. Wesman, A. G. "Writing the Test Item" in Educational Measurement, Second Edition, Robert L. Thorndike, ed. American Council on Education, Washington, D.C.; 1971, pp. 92-94.

4. Ebel, Robert L. "The Relation of Item Discrimination to Test Reliability"; Journal of Educational Measurement; 4:125-128, Fall, 1967.